A Section 508–conformant HTML version of this article is available at https://doi.org/10.1289/EHP1289.

A Web-Based System for Bayesian Benchmark Dose Estimation

Kan Shao[1] and Andrew J. Shapiro[2]

[1]Department of Environmental and Occupational Health, School of Public Health, Indiana University, Bloomington, Indiana, USA
[2]National Toxicology Program Division, National Institute of Environmental Health Sciences, National Institutes of Health, Department of Health and Human Services, Research Triangle Park, North Carolina, USA

Background: Benchmark dose (BMD) modeling is an important step in human health risk assessment and is used as the default approach to identify the point of departure for risk assessment. A probabilistic framework for dose–response assessment has been proposed and advocated by various institutions and organizations; therefore, a reliable tool is needed to provide distributional estimates for BMD and other important quantities in dose–response assessment.

Objectives: We developed an online system for Bayesian BMD (BBMD) estimation and compared results from this software with U.S. Environmental Protection Agency's (EPA's) Benchmark Dose Software (BMDS).

Methods: The system is built on a Bayesian framework featuring the application of Markov chain Monte Carlo (MCMC) sampling for model parameter estimation and BMD calculation, which makes the BBMD system fundamentally different from the currently prevailing BMD software packages. In addition to estimating the traditional BMDs for dichotomous and continuous data, the developed system is also capable of computing model-averaged BMD estimates.

Results: A total of 518 dichotomous and 108 continuous data sets extracted from the U.S. EPA's Integrated Risk Information System (IRIS) database (and similar databases) were used as testing data to compare the estimates from the BBMD and BMDS programs. The results suggest that the BBMD system may outperform the BMDS program in a number of aspects, including fewer failed BMD and BMDL calculations and estimates.

Conclusions: The BBMD system is a useful alternative tool for estimating BMD with additional functionalities for BMD analysis based on most recent research. Most importantly, the BBMD has the potential to incorporate prior information to make dose–response modeling more reliable and can provide distributional estimates for important quantities in dose–response assessment, which greatly facilitates the current trend for probabilistic risk assessment. https://doi.org/10.1289/EHP1289

Introduction

The benchmark dose (BMD) method has been widely accepted as the preferred method to replace the traditional no (or lowest) observed adverse effect level (NOAEL/LOAEL) approach for dose–response assessment in human health risk assessment. The BMD method has many important advantages over the NOAEL/LOAEL approach, but it requires more sophisticated regression algorithms to fit various dose–response models to the input data. Hence, it is necessary to have well-developed software to facilitate implementation of the BMD method.

There are two major software programs for BMD analysis that have been widely distributed and used by risk assessors and scientists throughout the world. The first is the Benchmark Dose Software [version 2.6.0.1; U.S. Environmental Protection Agency (EPA)] that was originally published by the U.S. EPA in 2000 and has been continuously upgraded and improved. This software is Windows based and has a well-designed graphical user interface (GUI) that is capable of analyzing multiple types of dose–response data, including the two most frequently used types: dichotomous data and continuous data. Over the years, a number of special dose–response models have been added to the software package (e.g., models to handle nested data) for certain specific uses, and some third-party packages (e.g., BMDS Wizard; ICF International) have been built to meet particular needs. The second software program, PROAST, is published by the Netherlands National Institute for Public Health and the Environment (RIVM). PROAST is programmed in the R programming language (R Core Team) and can be used on any operating system where R can be installed (e.g., Windows, Linux, Mac). PROAST is able to analyze dichotomous, continuous, and ordinal dose–response data, and a GUI was recently developed for the latest version, which was published in early 2014 (v.38.9). Both software packages have their respective advantages and are slightly different in some technical details, such as the dose–response models included and default assumptions on the distribution of continuous data. In general, both packages are suitable for dose–response analysis and deriving BMD and its statistical lower bound (BMDL). However, it is important to note that both software approaches utilize a frequentist-based statistical approach (i.e., the maximum likelihood estimation) for dose–response model fitting and parameter estimation.

In this paper, we present a web-based dose–response modeling system featuring an implementation of Bayesian inference for benchmark dose estimation. There are two important reasons for developing a Bayesian statistics–based BMD estimation system to supplement existing tools. First and most importantly, the Bayesian framework provides a way to incorporate prior information through the prior distribution of model parameters, which has great potential to enhance the reliability of dose–response modeling for poor-quality data, which may be the only data available for risk assessors in some situations. In addition, incorporating prior information may allow a reduction of the number of animals required for testing in future studies (Slob and Setzer 2014). Second, owing to the distributional/probabilistic nature of this approach, a Bayesian dose–response modeling tool can facilitate probabilistic risk assessment, which is advocated by the scientific community (Gaylor et al. 1999; Evans et al. 2001; Hattis et al.

Address correspondence to K. Shao, Department of Environmental and Occupational Health, Indiana University School of Public Health, 1025 E. Seventh Street, Bloomington, IN 47405 USA. Telephone: 812-856-2725. Email: kshao@indiana.edu

Supplemental Material is available online (https://doi.org/10.1289/EHP1289).

This article was co-written by Andrew J. Shapiro in his private capacity. No official support or endorsement by the NIH, National Institute of Environmental of Environmental Health is intended or should be inferred.

The authors declare they have no actual or potential competing financial interests.

Received 27 October 2016; Revised 18 November 2017; Accepted 21 November 2017; Published 11 January 2018; Corrected 23 March 2022.

2002; Axelrad et al. 2005; Woodruff et al. 2007; Chiu and Slob 2015). In 2009, the National Research Council (NRC) published a milestone work in the field of risk assessment, *Science and Decisions: Advancing Risk Assessment* (NRC 2009), which emphasizes the importance of probabilistically quantifying risk-specific dose in risk assessment to support regulatory decision making. More recently, the World Health Organization (WHO) published a guidance document on using probabilistic framework to harmonize the approaches to risk characterization (IPCS 2014).

The web-based application, which we refer to as the Bayesian Benchmark Dose (BBMD) analysis system, is available at https://benchmarkdose.org. Instead of providing point estimates for the quantities of interest (e.g., BMD, model parameters), the BBMD system characterizes a distribution of these quantities using Bayesian posterior samples. The software is designed to separate model fitting from BMD analysis; decoupling these steps makes the system computationally efficient and allows greater flexibility in analysis. The system implements recently developed BMD analysis approaches, such as the model-averaged (MA) BMD method (Shao and Gift 2014; Fang et al. 2015) and a hybrid approach for estimating BMD from continuous data (Crump 1995; Shao and Gift 2014). Thus, BBMD represents state-of-the-science methodology and technology in the field of dose–response assessment.

The paper is organized as follows: In the second section, a detailed introduction of the functionalities and features of the BBMD system is presented. In the third section, we compare BMD estimates from the BBMD system with their counterparts estimated from the U.S. EPA's BMDS (at present, the most widely used BMD estimation software). A comprehensive discussion of the advantages and limitations of the BBMD system is presented in the fourth section, followed by conclusions in the fifth section. Additional details on technical issues, test data sets, and BMD analysis results are provided in the Supplemental Material and in the "Testing Datasets and Results" file as indicated below.

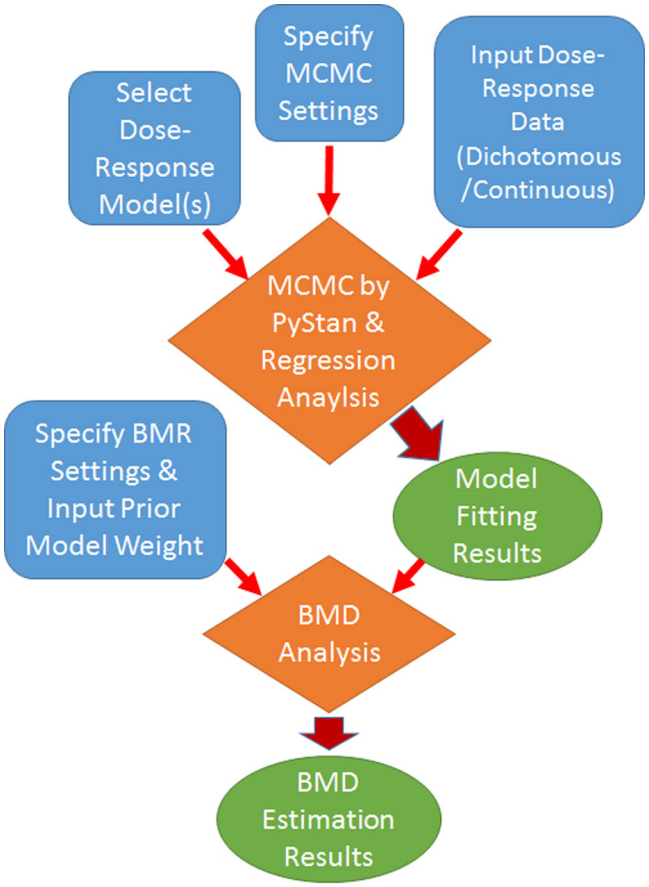


Figure 1. General steps to complete a benchmark dose (BMD) analysis in Bayesian Benchmark Dose (BBMD) system. Note: BMR, benchmark dose response; MCMC, Markov chain Monte Carlo.

The Bayesian Benchmark Dose (BBMD) Analysis System

Overview of the System

The BBMD system was developed to support quantitative dose–response assessment in human health risk assessment. The system was created using Python, C++, Stan, and Javascript programming languages and was designed as a web application. This particular online application has been published as open-source software with the Apache License (version 2.0), and the BBMD source code is available at https://github.com/kanshao. The system contains two modules: the back-end module, which is responsible for computation and data storage (see "Website Architecture" in the Supplemental Material for a more detailed description of the website architecture; see also Figures S1 and S2), and the front end module, which interacts with users via a web browser.

There are two views on the user interface: *a*) creating/updating an analysis and *b*) reviewing an existing analysis. In the first view, users create/change specific settings and execute the analysis; in the second view, analysis results can be displayed and exported. Figure 1 illustrates the general steps to complete a BMD analysis. Given an input dose–response data set, settings for the Markov chain Monte Carlo (MCMC) algorithm, and selected dose–response model(s), the system conducts model fitting; generates statistical estimates for model parameters, measures of goodness-of-fit, and cross-model comparison; and generates fitted dose–response curve(s). For BMD estimation, the posterior samples generated from the model-fitting process together with user-defined benchmark dose response (BMR) are further used in the final step for BMD estimation, in which graphical and textual results are presented to users. Detailed instructions on how to use the BBMD system are presented in the User Manual and Technical Guidance available on the BBMD website.

Input Data and Dose–Response Modeling Methods

Dichotomous and continuous dose–response data can be analyzed by BBMD. For each data type, the user can further specify input data as individual or summary data. The modeling strategy is the same for these different data types: It uses Bayesian inference to estimate model parameters based on Bayes' rule that can be expressed as

$$p(\theta|Data) \propto p(\theta)p(Data|\theta), \quad [1]$$

where $p(\theta|Data)$ and $p(\theta)$ are the posterior and prior distribution of the model parameters θ, respectively, and $p(Data|\theta)$ is the likelihood function. Equation 1 states that the posterior distribution of model parameters is proportional to the product of the prior distribution of model parameters and the likelihood function. For different data types, the likelihood function, $p(Data|\theta)$, differentiates in terms of distribution and model forms, which is discussed in detail below.

Dichotomous data. Dichotomous data use binary "success" or "failure" categories (1 or 0, respectively) to describe the status of subjects (e.g., animals tested in a toxicity study) treated at various dose levels with or without an effect (e.g., cancer). For summary dichotomous data, three variables [i.e., dose level (d_i), number of subjects in each dose group (n_i), and the number of

subjects with effect in the corresponding dose group (y_i)] are needed to characterize the dose–response relationship based on the assumption that the number of subjects with effect at each dose group follows a binomial distribution $y_i \sim binomial(n_i, r_i)$, where the parameter $r_i = f(d_i|\theta)$ represents the probability of effect in the ith dose group and is determined by a dose–response function with a parameter vector of θ (Shao and Small 2011). Given the settings of dichotomous data described above, the logarithm of the likelihood function in Equation 1 for summary dichotomous data can be expressed in Equation 2. The reason to express the likelihood function in logarithm format is that the log-likelihood function serves as the foundation for MCMC sampling in Stan (Stan Developmen Team).

$$\log[p(Data|\theta)] = \sum_{i=1}^{G}\left\{\log\binom{n_i}{y_i} + y_i \log[f(d_i|\theta)] + (n_i - y_i)\log[1 - f(d_i|\theta)]\right\}, \quad [2]$$

where G is the number of dose groups in the data set, and $f(d_i|\theta)$ represents a parametric dose–response model. There are eight choices available for the dose–response models for dichotomous data, and these models will be introduced in a later section.

For individual dichotomous data, each individual subject is characterized by two quantities: the dose level and a value of "1" or "0" indicating effect or effect-free. Random variables in this format can be described by a Bernoulli distribution. However, when considering the individual subjects exposed at the same dose level as a group, the number of subjects in each group (n) is fixed, and the number of subjects with effect (y) can be counted and dependent on probability (p), which is estimated by a dose–response function. Therefore, this data type is basically identical to the summary dichotomous data, which can be described by a binomial distribution. Consequently, for individual dichotomous data, the BBMD system will first convert the data to summary dichotomous data and then apply Equation 2 for modeling.

Continuous data. The second data type that can be modeled in BBMD software is continuous data (such as body weight and relative liver weight). For continuous data (regardless of whether they are individual or summary data), one fundamental assumption must be made regarding how the continuous responses are distributed. BMDS applies a normal distribution as the default assumption to model continuous data, and PROAST assumes that continuous responses are lognormally distributed. Shao et al. (2013) comprehensively examined and compared these two assumptions in the context of dose–response assessment and concluded that the lognormal assumption is more biologically plausible, adaptable, and reliable, particularly when within-group variance is large. Therefore, in the BBMD system, we use the lognormal distribution for modeling continuous end points.

If individual response data are available, the dose (d_i) and response (y_i) should be reported for each subject. Using θ' to represent parameters in a continuous dose–response model and γ^2 to represent the within-dose-group variance parameter (these two components together form the parameter vector θ in Equation 1), the log-likelihood function can be expressed as

$$\log[p(Data|\theta)] = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\gamma^2) - \frac{1}{2\gamma^2}\sum_{i=1}^{N}\left\{\log(y_i) - \log\left[f(d_i|\theta')\right]\right\}^2, \quad [3]$$

where $f(d_i|\theta')$ represents a dose–response model for continuous data, and N is the total number of subjects in the data set being analyzed. Available continuous dose–response in the BBMD system will be introduced in a later section.

In published literature, raw data are often unavailable, and results are reported as summary statistics (e.g., mean and standard deviation), which we refer to as summary continuous data in BBMD. In a complete summary continuous data set, there are four reported variables: dose level (d_i), number of subjects in each group (n_i), the observed mean value of response in each group ($\bar{y}_i$), and the observed standard deviation of response in each group (s_i). Under the lognormality assumption, the commonly reported mean and standard deviation on the regular scale are converted to the corresponding quantities on a log scale using $\bar{y}_i' = \log(\bar{y}_i) - 0.5 \times \log[(s_i/\bar{y}_i)^2 + 1]$ and $s_i' = \sqrt{\log[(s_i/\bar{y}_i)^2 + 1]}$ (Crump 1995, Slob 2002). Modeling summary continuous data shares the same fundamental idea as Equation 3 but is differentiated by input data, so the log-likelihood function that is used in the Stan model for MCMC sampling is

$$\log[p(Data|\theta)] = -\frac{N}{2}\log(2\pi) - \sum_{i=1}^{G}\left\{\frac{n_i}{2}\log(\gamma^2) + \frac{n_i \times \left\{\bar{y}_i' - \log[f(d_i|\theta')]\right\}^2 + (n_i - 1) \times s_i'^2}{2\gamma^2}\right\}, \quad [4]$$

where N and G represent the total number of subjects and the number of dose groups, respectively.

It is worth mentioning that the U.S. EPA's BMDS allows the user to input individual continuous data, but the system converts the individual data into summary data by grouping subjects with the same dose level and then fits the data as if they were summary data. For many toxicological bioassay and *in vitro* data sets, this is a valid approach given the normality assumption used in the BMDS because multiple animals/replicates are usually tested at the same dose level. However, this approach limits the capability of analyzing responses with unique dose levels (i.e., $n = 1$ for each dose group, a common situation in epidemiological data sets) because one required input quantity [s_i' (or s_i, the within-dose-group standard deviation)] in Equation 4 cannot be calculated when $n = 1$. Hence, the ability of the BBMD system to directly model individual continuous data is important.

Modeling Settings

The modeling settings consist of two major components: the Markov chain Monte Carlo (MCMC) settings and dose–response model settings.

Markov chain Monte Carlo (MCMC) settings. MCMC settings provide specifications to Stan (Carpenter et al. 2017) where the posterior sampling via MCMC is performed. Users specify the length of the Markov chain (i.e., how many samples to generate in each chain), the number of chains, the warm-up ratio (the proportion of posterior sample in each chain to be discarded), and a random seed. The chain length and warm-up ratio are closely related to the posterior sample convergence, an important indicator to judge MCMC sampling performance. In principle, the longer the chain, the better the chance that the chain will eventually converge, so using longer chains and a larger warm-up ratio is a way to ensure convergence. Because multiple chains may use different sets of initial values, these chains that converge to the same steady state can further justify the reliability of the sampling algorithm. However, longer chains require greater computational time and increased data storage requirements, which is a critical issue for an online system. Based on testing and

calibration of the BBMD modeling system, Stan generally appears to converge relatively quickly on most data sets tested, so it is often unnecessary to use longer chains. The default settings used by the BBMD system (one chain with 30,000 samples and 50% warm-up ratio) perform well based on testing results (presented in a later section). The fourth MCMC setting, random seed, is randomly generated for each analysis but can also be specified by the user if a fixed seed is desired (for result reproducibility).

Dose–response model settings. Sixteen frequently used dose–response models are presently available in BBMD, eight for dichotomous data and eight for continuous data. The models with parameter value ranges are listed in Appendix 1.

There is a "g" parameter (a power parameter on the dose) in the Weibull, Loglogistic, Logprobit, and Dichotomous Hill models for dichotomous data, and in the Power, Hill, Exponential 3, and Exponential 5 models for continuous data. According to the BMD Technical Guidance (U.S. EPA 2012), the default setting in the U.S. EPA's BMDS for this power parameter is ≥ 1, with the option to relax the restriction to ≥ 0. There is debate in the scientific community about whether the power parameter should be restricted to ≥ 1. In the present BBMD system, we provide five options for the power parameter in the corresponding models: ≥ 0, ≥ 0.25, ≥ 0.5, ≥ 0.75, and ≥ 1.

Prior distributions for model parameters. The prior distribution is one important component in a Bayesian framework. In the present BBMD system, uniform distributions are used for all model parameters, and the lower and upper bounds of these uniform distributions are determined based on biological considerations and preliminary testing. The specification of priors needs to balance the flexibility of the model and the unnecessary uncertainty in estimation, so the range of the parameters determined by the uniform prior distribution cannot be too large or too small. (See "Settings of Prior for Model Parameters" in the Supplemental Material for details on the uniform distributions used for different models). Testing results presented and discussed below detail the appropriateness of using these prior distributions.

Instead of using noninformative priors (e.g., the uniform distributions employed in the present BBMD system), properly derived informative priors may enhance the reliability of model fitting and BMD estimation. However, effectively employing informative priors requires extensive research; this will be our next major task in BBMD development. (See "The Impact of Generalized Informative Prior on BMD Estimation" in the Supplemental Material for a preliminary example demonstrating how using informative prior may impact BMD estimation; see also Figures S3–S6).

Dose–Response Modeling Results

After a data set and settings are provided, the BBMD system performs regression analysis and provides outputs in the "Model fit results" tab. The following statistics are available for each model: model parameter estimates, posterior predictive *p*-value, model weight, and graphical dose–response curve.

Model parameter estimation. The first section on the "Model fit results" page contains summary statistics (including mean, standard error of the mean, standard deviation, various quantiles, and quantities that indicate effective sample size and chain convergence) for each model parameter. The data shown in the text box in the upper part of Figure 2 are directly acquired from the Stan output; some information regarding the MCMC execution is also presented. A correlation coefficient matrix is provided for the model parameters and is calculated using posterior samples.

For the purpose of generality, doses in all data sets are normalized to the scale between 0 and 1 by dividing the highest dose level in that data set. Therefore, to reproduce dose–response curves or to calculate BMDs, parameter summary statistics shown in the box or the posterior samples of model parameters exported from the website should not be directly employed for such activities. Instead, maximum dose level in this particular data set needs to be properly applied for plotting the dose–response curve or calculating the BMD.

An additional graphical output can be expanded by the user at the bottom of this page to show the probability density plot estimated by the kernel density estimation function in SciPy (Oliphant 2007; Millman and Aivazis 2011) and the trace plot of posterior samples for each model parameter, shown in Figure 3 as an example. Being cognizant of a user's web browsing experience, parameter chain plots are hidden by default to avoid unnecessary transmission of large data sets.

Posterior predictive p-value. Posterior predictive *p*-value (PPP; Gelman et al. 2004) is a way to assess the fit of the model to the data under the Bayesian framework and has a similar purpose as the *p*-value provided in traditional BMD software, such as BMDS, which uses frequentist statistical approaches. However, the *p*-values are interpreted differently. Both *p*-values use likelihood as a key statistic. In the U.S. EPA's BMDS, a likelihood ratio (the likelihood of the fitted model over the likelihood of the saturated model) is assumed following a χ^2 distribution, and the null hypothesis is rejected if the *p*-value is too small (i.e., model fitting is not adequate). In practice, as recommended in the BMD Technical Guidance (U.S. EPA 2012), if the *p*-value is <0.1, the model should not be considered for BMD calculation. The BBMD system uses the posterior samples of model parameters and estimates the posterior predictive *p*-value as described by Gelman et al. (2004). Posterior samples are first used to generate predicted responses; then, likelihood values calculated using the predicted responses and original data are computed and compared; finally, the probability that one type of likelihood is larger than the other (e.g., $\Pr[T(y,\theta^l) > T(y^{pred,l},\theta^l)]$) is estimated. The PPP can be approximated by counting the predicted responses that satisfy the inequality out of the entire posterior sample space. A large or small *p*-value means that a discrepancy in predicted data is very likely, further indicating a poor fit. Therefore, a PPP value within the range from 0.05 to 0.95 indicates an adequate fit.

Model weight calculation. A model weight is calculated for each model included in the analysis as a statistic for cross-model comparison. The approach used in the system to compute model weight was introduced by Wasserman (2000), using the following two equations [i.e., Equations 25 and 26 in Wasserman (2000)]:

$$\Pr(\mathcal{M}_j|Data) = \frac{\widehat{m}_j}{\sum_{k=1}^{K} \widehat{m}_k} \quad [5]$$

and

$$\log(\widehat{m}_j) = \widehat{\ell}_j - \frac{q_j}{2}\log(n), \quad [6]$$

where $\widehat{\ell}_j = \ell_j(\widehat{\theta}_j)$ is a log-likelihood value estimated using posterior sample of model parameters of the *j*th model; q_j is the number of model parameters in the *j*th model, and n is the data set sample size. The meaning of Equation 5 is that the posterior model weight of model j is equal to the m value estimated from model j divided by the sum of m values estimated from all models in the analysis. Equation 5 is a special case of Equation 13 (defined below and used for MA BMD estimation), when all models in the analysis have an equal prior weight (i.e., $1/K$). The m value for each model is calculated using Equation 6; the right side of the equation is similar to the Bayesian Information Criterion (BIC) model weight approximation method originally proposed by Kass and Raftery (1995) and widely applied in more

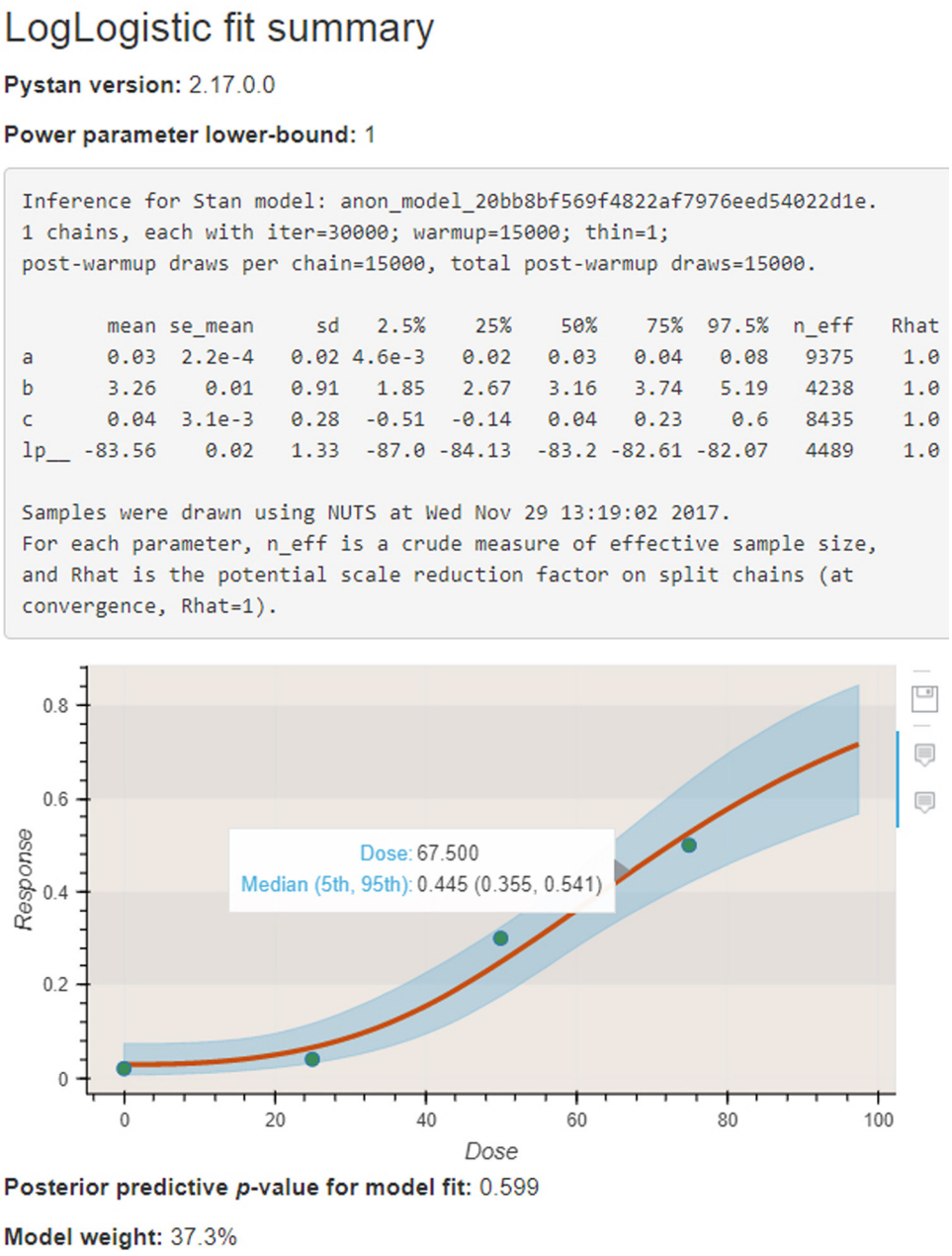


Figure 2. Textual and graphical output for model fitting results. The textual output in the box mainly includes the mean, standard error of the mean, standard deviation, various quantiles, and quantities indicating effective sample size and chain convergence for each model parameter, as well as information regarding the Markov chain Monte Carlo (MCMC) execution. A dynamic dose–response plot is shown below the text box. This plot includes original dose–response data and a fitted curve with its 90th percentile interval shaded in blue. The estimated median and the 5th and and 95th percentiles at a particular dose level indicated by the user's cursor are also displayed. Other information displayed in this figure includes the PyStan version, the lower bound placed on the power parameter (if applicable), the posterior predictive *p*-value (PPP value) for model fit and model weight for cross-model comparison.

recent dose–response assessment literature (Wheeler and Bailer 2007; Shao and Gift 2014). Each set of the posterior sample of model parameters is used to calculate a set of posterior model weights for the models included. The reported model weights are the average posterior weights of each model. Although the model weights calculated are based on likelihood (with no preference to model format) and are used for cross-model comparison, this approach provides a base for model averaging to address model uncertainty, further discussed below in "Model-averaged BMD calculation."

In the BMDS software, the Akaike Information Criterion (AIC) value is computed and is used for comparing fitted dose–response models. The AIC, like the BIC, includes both likelihood and a penalty term for the number of parameters. However, the AIC mainly provides a qualitative model comparison (i.e., a model with a lower AIC value is better, but how much better is difficult to discern). The model weight approach implemented in BBMD provides numerical model weights for each model, which is advantageous in the context of probabilistic risk assessment to the AIC method because the weights quantitatively compare dose–response models and probabilistically quantify model uncertainty.

An implicit assumption in Equation 5 is that each model included in the analysis has an equal prior model weight, which means that we believe each model is equally likely to fit the data well before we see the data. This assumption makes the model weights reported on this page solely determined by model fit and

Correlation matrix:

	a	b	c
a	1	0.312	-0.181
b	0.312	1	0.416
c	-0.181	0.416	1

Parameter charts:

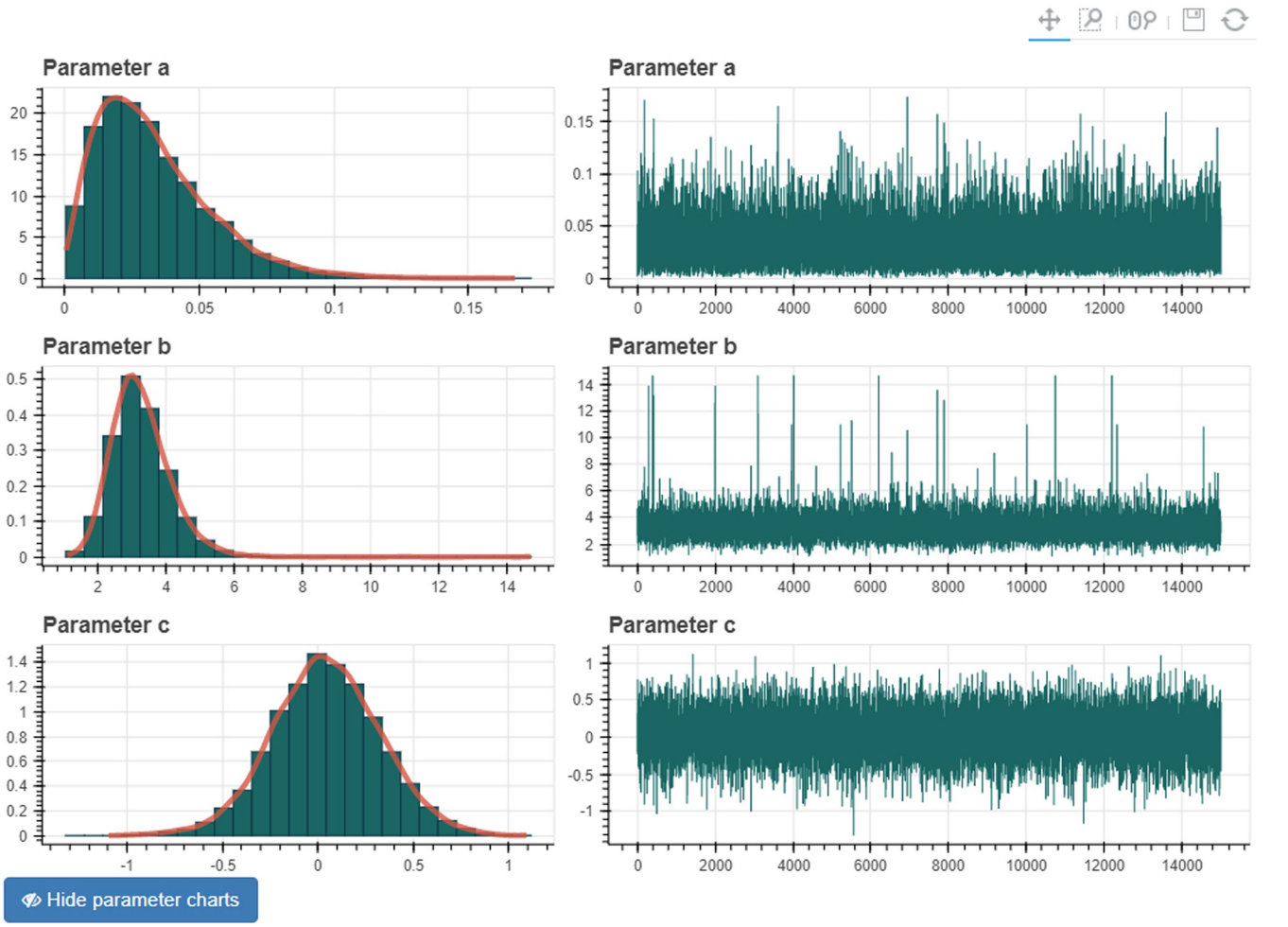


Figure 3. Distribution plot and posterior sample tracing plot for model parameters. This figure is a screenshot of the Bayesian Benchmark Dose (BBMD) website. A correlation matrix is displayed on the top of the graph to show the correlation coefficients between different model parameters. On the bottom, a distribution plot (including a histogram and fitted probability density curve) and a posterior sample tracing plot (i.e., all posterior samples are sequentially connected by solid lines) are illustrated for each parameter.

the number of parameters in the models. In the BMD estimation step described in "Model-averaged BMD calculation," the model weight concept is used again, but additional information on prior model weight will be incorporated.

The dose–response plot. The lower part of Figure 2 shows an example of the dynamic dose–response plot presented on this tab. The plot displays the original data and the fitted curve (solid orange line) with its 90th percentile interval (shaded area in blue). The plot also interactively displays the estimated median and the 5th and 95th percentiles of the response at the dose level where a user moves the mouse. This feature allows users to capture the response range at doses of interest (or vice versa).

Benchmark Dose Estimation

A benchmark dose is defined as the dose level that causes a predetermined change in the response, and it has a number of uses in risk assessment applications. The BBMD software provides distributional BMD estimates using multiple definitions and derivations. In addition, recently developed model-averaging methodology has been implemented for estimating MA BMD, which may be a preferred method for BMD analysis (Wheeler and Bailer 2007; Shao and Gift 2014; EFSA Scientific Committee et al. 2017) because it represents an ensemble of all models used in the analysis.

BMD calculation for dichotomous data. For dichotomous data, two commonly used BMD definitions are extra risk and added risk, as shown in Equations 7 and 8, respectively:

$$BMR = \frac{f(BMD) - f(0)}{1 - f(0)} \quad [7]$$

and

$$BMR = f(BMD) - f(0), \quad [8]$$

where $f(\cdot)$ represents a dichotomous dose–response model. *BMR* in Equations 7 and 8 stands for benchmark response, which is a specified increase in the probability of response and is commonly set at 10%, 5%, or 1%. BMD based on extra risk definition is the default option used in BMDS, and only one BMR can be selected for each model execution. The BBMD system calculates both BMDs for each dichotomous dose–response model included in an analysis.

Under a Bayesian framework, BMD estimation is basically calculating the posterior sample of BMD with the same length as the posterior sample of the model parameters. With the posterior sample, a number of statistics (including the mean, median, standard deviation, and other quantiles) of BMD can be computed and are reported on the "BMD estimates" tab. Based on our testing, the median value of the BMD posterior sample is the most reliable estimate for BMD owing to its resistance to some extreme values in the sample. In addition, the 5th percentile of the posterior sample is considered the lower bound of BMD (i.e., BMDL) corresponding to the lower bound of the one-sided 95th confidence interval used in the U.S. EPA's BMDS. The BMDL is usually used as the point of departure for low-dose extrapolation and is therefore of great regulatory interest. The same procedures used for determining BMD and BMDL are also applied to continuous data.

BMD calculation for continuous data. For continuous data, multiple BMD definitions are available in BBMD and are grouped into two categories: *a*) based on central tendency and *b*) based on tails [i.e., the hybrid approach proposed by Crump (1995)].

For BMD defined on central tendency, there are three options for defining the BMR value: *a*) relative change, *b*) absolute change, and *c*) cutoff, which are expressed by Equations 9–11, respectively:

$$f(BMD) \pm f(0) = Relative\ Change \times f(0), \quad [9]$$

$$f(BMD) = Absolute\ Change \pm f(0), \quad [10]$$

$$f(BMD) = cutoff, \quad [11]$$

where $f(\cdot)$ represents a continuous dose–response model fit to the central tendency of the data (i.e., the median under the lognormality assumption). The BMD is the dose level that satisfies the selected definition equation. For continuous data, both increasing and decreasing trends are possible, so there is a "$\pm$" in Equations 9 and 10 corresponding to increasing or decreasing trend.

For a BMD defined using a hybrid approach, an adversity value must be specified in addition to a BMR value. The hybrid approach considers any response above or below (i.e., corresponding to increasing or decreasing trend) the adversity value as abnormal; thus, the BMD is the dose level where the proportion of the abnormality has increased a certain percent (i.e., BMR) compared with the control. Mathematically, for increasing trend, the hybrid BMD definition can be expressed as $Q(0) - Q(BMD) = BMR$ for added risk, and $\left[Q(0) - Q(BMD)\right]/\left[1 - Q(0)\right] = BMR$ for extra risk, where $Q(\cdot)$ is the quantile of the adversity cutoff value at a specified dose level.

It remains debatable whether the hybrid method is superior or more biologically plausible than central tendency methods. However, in addition to the original publication (Crump 1995), recent publications (Shao and Gift 2014; Wheeler et al. 2015, 2017) have accepted the idea of defining the BMD based on the tails of the distribution. Therefore, we believe the hybrid method can provide a useful supplemental approach for BMD estimation using continuous data. The BBMD software is the first benchmark dose software with a graphical user interface that implements the hybrid approach.

Model-averaged BMD calculation. BBMD allows for the calculation of MA BMDs and BMDLs, which have been recommended for use by the European Food Safety Authority (EFSA Scientific Committee et al. 2017). Based on the idea of model averaging introduced by Hoeting et al. (1999), the MA BMD can be expressed as

$$\Pr(BMD_{ma}|Data) = \sum_{k=1}^{K} \Pr(BMD_k|\mathcal{M}_k, Data)\Pr(\mathcal{M}_k|Data). \quad [12]$$

The explanation of Equation 12 is that the MA distribution of the BMD is a weighted sum of the BMD distribution estimated from each individual model included in an analysis. The model weight portion of Equation 12, $\Pr(\mathcal{M}_k|Data)$, was previously described in Equation 5 for cross-model comparison (Equation 5 assumes equal model prior weights). In a Bayesian context, the distribution of BMD is characterized by a vector of posterior sample of BMD. Hence, the MA BMD vector is an integration of weighted vectors from individual models. Then, the vector of MA BMDs (which is the same length as an individual model posterior sample) can be used to compute statistics such as the BMD_{ma} (the median) or the $BMDL_{ma}$ (the 5th percentile). A few different methods have been proposed for MA BMD calculations (Wheeler and Bailer 2007; Shao and Gift 2014; Simmons et al. 2015; Fang et al. 2015), and simulation study is needed to judge which is superior or whether these methods are generally similar. The primary reason for selecting the method described above [similar to the method proposed by Shao and Gift (2014)] is that this method is effective and consistent with the Bayesian modeling method employed in the BBMD system.

The model weight calculation equation is now expanded from Equation 5 by adding a prior model weight as expressed by the equation below [i.e., Equation 18 in Wasserman (2000)]:

$$\Pr(\mathcal{M}_j|Data) = \frac{\widehat{m}_j \Pr(\mathcal{M}_j)}{\sum_{k=1}^{K} \widehat{m}_k \Pr(\mathcal{M}_k)}. \quad [13]$$

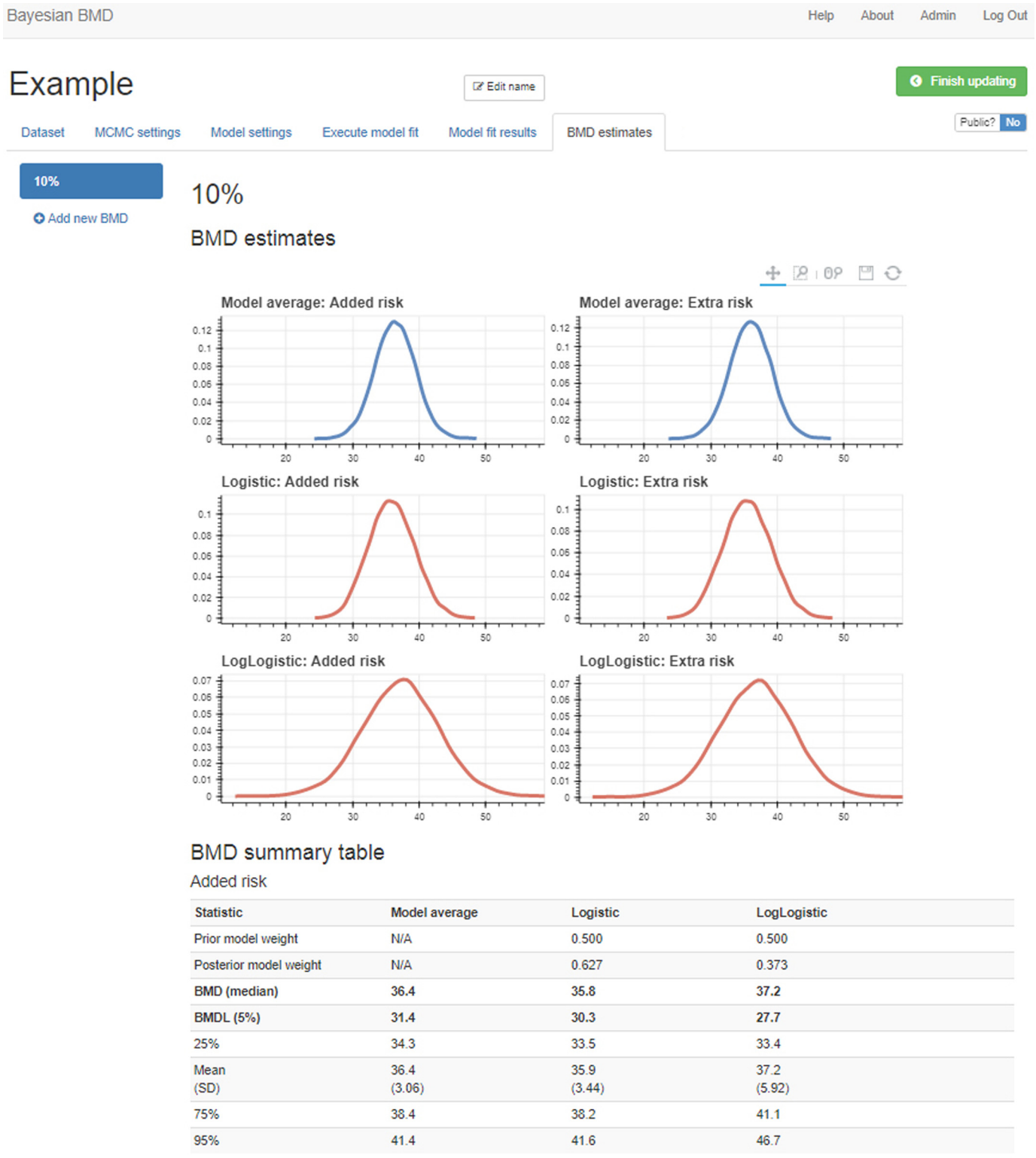
The prior weight of individual models, $\Pr(\mathcal{M}_j)$, has a default value of $1/K$ (where K is the number of models included in an analysis), but each model's prior weight can be modified in BBMD. The ability to modify the model prior weight serves two primary functions:

- First, it allows users to include prior knowledge into the BMD analysis. For example, if previous analyses or collective toxicological knowledge suggest that one model is preferred over others, a higher prior weight can be given to the preferred model.
- Second, it gives users a second chance to include/exclude models for the MA BMD calculation. For example, a user may choose to use the Weibull model twice in the model-fitting step, using different settings for the power parameter (e.g., ≥ 0 and ≥ 1). After the model fitting, instead of using both Weibull models in MA BMD calculations, the user can exclude one Weibull model instance by specifying a prior model weight of 0.

These features make the BBMD system unique and advanced compared with the existing BMD software packages; they also represent the state-of-the-science technology in dose–response modeling. A screen shot of the BMD estimation page in the BBMD system is shown in Figure 4.

Testing Results Comparison

In an effort to better understand the results from BBMD, we tested the system by comparing outputs from BBMD and from



BMD summary table

Added risk

Statistic	Model average	Logistic	LogLogistic
Prior model weight	N/A	0.500	0.500
Posterior model weight	N/A	0.627	0.373
BMD (median)	**36.4**	**35.8**	**37.2**
BMDL (5%)	**31.4**	**30.3**	**27.7**
25%	34.3	33.5	33.4
Mean (SD)	36.4 (3.06)	35.9 (3.44)	37.2 (5.92)
75%	38.4	38.2	41.1
95%	41.4	41.6	46.7

Figure 4. Graphical and tabular output for benchmark dose (BMD) estimates. This example presents two dichotomous dose–response models, Logistic and Loglogistic, along with a single 10% benchmark dose response (BMR), shown in the form of both Extra risk and Added risk. The model average of both models is also present. The figures present the probability distribution function (PDF) of BMD estimates for each model. The table below presents the prior model weights, the posterior model weight, and various statistics for each individual model and the model average.

the U.S. EPA's BMDS software, modeling the same data sets using the default parameter settings (i.e., power parameters are restricted to be ≥ 1). The data sets were actual toxicological data reported in the U.S. EPA Integrated Risk Information System (IRIS) toxicological review reports, as well as other similar reports by the U.S. EPA and other agencies. Wignall et al. (2014) proposed a standardized procedure to estimate BMD/BMDL using BMDS and applied the procedure to these data sets.

Together with the manuscript, the authors published the data sets used in their study and suggested actions to be taken (e.g., excluding the highest dose level in the data set for BMD analysis). We used all data sets that were suitable for BMD analysis (i.e., at least one of the dose–response models could adequately fit the data for BMD calculation) and removed duplicates, leaving 518 dichotomous data sets and 108 continuous data sets for testing and model comparison.

Our analysis included a model-wise comparison of quantities of interest that can be used to judge the quality and reliability of BMD estimation. The BMDS and BBMD systems fit the same dose–response models to the same data sets and estimated BMDs and BMDLs based on various BMR definitions. For dichotomous data, both extra-risk and added-risk BMD definitions were tested at two BMR values (10% and 1%): a total of four combinations (i.e., $518 \times 4 = 2{,}072$ BMDs and 2,072 BMDLs were calculated for each model). For continuous data, owing to the difference in the assumption on the distribution of responses used in the two systems (normal in BMDS and lognormal in BBMD), BMDs defined by relative change in central tendency were considered sufficiently comparable (Shao et al. 2013). Thus, we compared the BMDs (and associated BMDLs) estimated using 10% and 1% relative change in the central tendency (i.e., $108 \times 2 = 216$ BMDs and 216 BMDLs were calculated for each of the seven continuous models included in both software packages). Polynomial is not included in BBMD; Michaelis-Menten is not included in BMDS). The following common and program-specific comparison quantities are measured for each dose–response model and are presented in Tables 1 and 2 for dichotomous and continuous data, respectively:

1. Number of failed BMD, a common measure between software packages. "Failure" is defined as the BMD estimates being reported as "not available (NA)" or "error" or ≤ 0. There are 2,072 BMD estimates for dichotomous data and 216 BMD estimates for continuous data.
2. Number of failed BMDL, a common measure, where "failure" is defined as above in (1).
3. The BMD/BMDL ratio, a common measure. Models with either a failed BMD or BMDL as defined above in (1) and (2) were removed from this analysis. Because the BMD/BMDL ratio is commonly used for measuring the reliability of BMDL estimates regardless of BMR definition, the extra-risk BMDs and added-risk BMDs are considered together. However, model performance may change dramatically in low dose ranges; thus, the ratios at the BMR = 10% and 1% (and the relative change at 10% and 1% for continuous data) are calculated separately. The median and the 95th percentile interval of the ratio were reported.
4. Number of reduced models, a BMDS-specific measure. In BMDS, a more complex model may reduce to a simpler form if one or more parameters hit the defined parameter bound during the optimization process. For example, the Weibull model will become the quantal-linear model when the power parameter hits the bound at 1. When the model format is simplified, BMDL estimation may be affected (because the number of parameters is reduced). Therefore, in this analysis, if a more complex model has AIC and model fitting p-values (simultaneously) identical to those of a plausibly simpler model (for example, the quantal-linear model is a plausibly simpler model for the second-degree multistage model but not for the LogProbit model), then the complex model is considered as "reduced". This indicator is model-specific and is not available for all models in BMDS (e.g., a two-parameter model cannot be simplified further). In BBMD, posterior sampling for all model parameters is

Table 1. Comparison of BMD estimation for dichotomous data.

Quantities measured	Quantal-linear	Logistic	Probit	Weibull	Multistage 2	LogLogistic	LogProbit	Dichotomous Hill
BMDS								
Number of failed BMD	0	0	0	12	0	0	4	773[a]
Number of failed BMDL	0	8	0	12	1	0	8	833[a]
BMD/BMDL ratio	1.51	1.30	1.31	1.70	1.62	1.89	1.49	1.69
(at BMR = 0.1)	(1.21–2.69)	(1.13–3.19)	(1.15–3.03)	(1.20–8.41)	(1.18–5.73)	(1.21–10.5)	(1.20–4.75)	(1.11–10.3)
BMD/BMDL ratio	1.51	1.50	1.51	2.51	2.14	3.22	1.65	4.91
(at BMR = 0.01)	(1.21–2.67)	(1.22–15.5)	(1.20–13.9)	(1.24–56.2)	(1.24–18.6)	(1.42–68.0)	(1.24–10.2)	(1.23–93.6)
Number of reduced model	NA	NA	NA	183 to Quantal-linear	184 to Quantal-linear	31 to Logistic	63 to Probit	124 to LogLotistic
BBMD								
Number of failed BMD	0	0	0	0	0	0	0	0
Number of failed BMDL	0	0	0	0	0	0	0	0
BMD/BMDL ratio	1.53	1.29	1.29	1.69	1.60	1.77	1.47	2.31
(at BMR = 0.1)	(1.21–2.51)	(1.09–2.20)	(1.10–2.06)	(1.12–4.39)	(1.24–2.59)	(1.13–5.40)	(1.08–3.81)	(1.19–190.7)[b]
BMD/BMDL ratio	1.53	1.51	1.50	3.38	2.23	3.56	2.00	4.23
(at BMR = 0.01)	(1.21–2.50)	(1.22–4.30)	(1.20–3.92)	(1.42–17.5)	(1.31–3.49)	(1.51–19.36)	(1.28–7.01)	(1.35–593)[b]
Comparison								
Correlation coefficient for BMD	0.991	0.998	0.997	0.842	0.969	0.830	0.857	0.837
Correlation coefficient for BMDL	1.000	0.985	0.978	0.945	0.988	0.898	0.955	0.855
Ratio of BMDs	1.00	1.02	1.02	1.57	0.929	1.54	1.58	1.26
	(0.829–1.18)	(0.714–1.25)	(0.494–1.32)	(0.481–24.7)	(0.205–1.67)	(0.737–29.8)	(0.865–8.98)	(0.530–29.8)
Ratio of BMDLs	1.00	1.03	1.02	1.68	1.06	1.93	1.66	1.59
	(0.888–1.89)	(0.973–2.44)	(0.942–2.71)	(1.02–9.63)	(0.530–1.29)	(1.05–18.0)	(1.06–6.10)	(0.079–21.5)

Note: BBMD, Bayesian benchmark dose method; BMD, benchmark dose; BMDL, lower bound of BMD; BMR, benchmark response; BMDS, U.S. Environmental Protection Agency's Benchmark Dose Software; NA, not available.

[a]The BMDS directly reports "error" for BMD and BMDL when the number of dose groups is smaller than the number of model parameters in the Dichotomous Hill model. Of the 518 data sets, 186 have only three dose groups; therefore, 744 ($= 186 \times 4$) in these failed BMDs or BMDLs are due to insufficient dose groups.

[b]For the BMD/BMDL ratios calculated using the Dichotomous Hill model in the BBMD system, all results from the 518 data sets (including those having only three dose groups) are included.

Table 2. Comparison of BMD estimation for continuous data.

Quantities measured	Linear	Power	Hill	Exponential 2	Exponential 3	Exponential 4	Exponential 5
BMDS							
Number of failed BMD	2	0	34[a]	0	0	2	36[a]
Number of failed BMDL	2	2	38[a]	1	1	3	37[a]
BMD/BMDL ratio	1.28	1.39	2.16	1.28	1.34	1.54	2.16
(at relative change = 0.1)	(1.07–2.85)	(1.05–12.9)	(1.08–1.72×10^7)	(1.07–2.14)	(1.07–6.97)	(1.09–207)	(1.13–441)
BMD/BMDL ratio	1.28	1.85	4.49	1.27	1.63	1.65	4.64
(at relative change = 0.01)	(1.07–2.85)	(1.07–33.4)	(1.20–1.32×10^6)	(1.07–2.14)	(1.07–46.96)	(1.11–211)	(1.32–985)
Number of reduced model	NA	52 to Linear	NA	NA	57 to Exponential 2	24 to Exponential 2	22 to Exponential 3/4
BBMD							
Number of failed BMD	0	0	1	0	0	0	0
Number of failed BMDL	0	0	1	0	0	0	0
BMD/BMDL ratio	1.27	1.33	2.05	1.25	1.30	1.59	1.98
(at relative change = 0.1)	(1.07–2.28)	(1.06–4.50)	(1.12–11.3)[b]	(1.07–2.16)	(1.06–5.66)	(1.17–22.5)	(1.06–32.5)[b]
BMD/BMDL ratio	1.27	3.07	3.91	1.25	3.29	1.69	3.95
(at relative change = 0.01)	(1.07–2.28)	(1.13–23.0)	(1.44–36.1)[b]	(1.07–2.16)	(1.12–25.1)	(1.22–19.6)	(1.44–25.8)[b]
Comparison							
Correlation coefficient for BMD	0.999	0.946	0.822	0.989	0.919	0.960	0.805
Correlation coefficient for BMDL	0.994	0.960	0.927	0.992	0.950	0.861	0.847
Ratio of BMDs	0.988	1.22	1.13	0.988	1.34	0.874	1.05
	(0.685–1.29)	(0.797–34.0)	(0.036–1,537)	(0.823–1.27)	(0.848–32.8)	(0.113–1.32)	(0.093–7.57)
Ratio of BMDLs	0.994	1.43	1.68	0.986	1.41	0.871	1.30
	(0.719–2.09)	(0.916–10.0)	(0.639–4.5×10^6)	(0.802–1.37)	(0.954–11.7)	(0.039–94.3)	(0.080–181)

Note: BBMD, Bayesian benchmark dose method; BMD, benchmark dose; BMDL, lower bound of BMD; BMDS, U.S. Environmental Protection Agency's Benchmark Dose Software; NA, not available.
[a]The BMDS directly reports "error" for BMD and BMDL when the number of dose groups is smaller than the number of model parameters in the Hill and Exponential 5 models. Of the 108 data sets, 16 have only three dose groups; therefore, 32 ($=16\times2$) in these failed BMDs or BMDLs are due to insufficient dose groups.
[b]For the BMD/BMDL ratios calculated using the Hill and Exponential 5 models in the BBMD system, all results from the 108 data sets (including those having only three dose groups) are included.

always performed; therefore, models cannot reduce to a simpler form, even with restricted parameters.

5. Comparison between the BBMD and BMDS systems. The comparison focuses on two measures: the correlation coefficient and the ratio of BMD and BMDL estimates obtained from the BBMD and BMDS estimates, respectively.
   a. The Pearson's correlation coefficient (PCC) was calculated for "failure-removed" BMD estimates from the BBMD system (variable 1) and from the BMDS system (variable 2) where the BMDs based on different definitions were not separated for this calculation because this value should not be BMR-type-specific. The PCC was also calculated for BMDL estimates.
   b. The ratio of the BMD and BMDL estimates from the two systems are calculated and reported in median and 95th percentile intervals in Tables 1 and 2. In these two tables, the ratios are not calculated separately for different BMD definitions or BMR values. In addition, to compare the BMD and BMDL estimates from the two systems, we used linear models to fit the BMD (or BMDL) estimates and generate BMD-BMD or BMDL-BMDL plots (plots and estimated coefficients are provided in Figures S7–S22 for dichotomous data and in Figures S23–S36 for continuous data).

In BBMD, convergence of posterior sampling is an important consideration because it indicates the reliability and consistency of the MCMC posterior sampling, which is important for characterizing BMD distributions. Convergence is measured by $\widehat{R}$ for each parameter distribution in Stan, which is subsequently reported in BBMD. The closer to 1 the value of $\widehat{R}$ is, the better convergence is achieved. In the testing analyses, we first used the default settings (e.g., 1 chain, 30,000 samples in length, the first 15,000 samples treated as warm-up and not saved in the posterior) to analyze all data sets. To be conservative, the maximum $\widehat{R}$ value among all parameters in a model was reported as the $\widehat{R}$ for that model. Based on the results of the first round, model/data set combinations that had a weak convergence measure ($1.01 \le \widehat{R} < 1.05$) were identified, and the length of the MCMC chains was increased in the second round to $30{,}000 \times 2 = 60{,}000$. For model/data set combinations that had a very weak convergence measure ($\widehat{R} \ge 1.05$), the length of the chain was changed to 120,000. For these two customized situations, the final posterior MCMC sample to be used in analyses was kept at 15,000 for each model; thus, the length of the warm-up sample may vary (i.e., the percentage of the warm-up sample is 75% and 87.5% in these two situations, respectively). For the model/data set combination that was well-converged in the first round, the default setting was kept. The percentage of data sets with $\widehat{R} \le 1.01$ and the 97.5th percentile of the $\widehat{R}$ value for each dose–response model are reported in Table 3 for the first- and second-round analyses. We also tried increasing the number of chains and the length of each chain for poorly converged data set/models, but we found that the convergence performance was not better than the customized strategy reported in Table 3 (for some models, the performance was even worse). Therefore, these results suggest that the default MCMC settings used in the BBMD system are adequate for most data sets (assuming that future data sets are similar to the real toxicological data used in testing).

Another important measure provided in BBMD is effective sample size, which gives a sense of whether the simulated sample is sufficient for practical purposes. In Table 3, the mean and the 95th percentile interval of the minimum effective sample size (i.e., the smallest effective sample size among model parameters was chosen as the effective sample size for that data set) estimated from all testing data sets are presented for each model. The results suggest that the default settings used in the BBMD system can provide adequate effective sample size.

Additionally, we examined some important sampler parameters to determine the quality of the MCMC sampling, mainly including

Table 3. Analytics on the of effective sample size and $\widehat{R}$ (indicating the convergence of MCMC sampling).

Model	Minimum effective sample size, default setting Mean (95% CI)	Default setting		Customized MCMC length	
		Percent of data set with $\widehat{R} \leq 1.01$	97.5th percentile of $\widehat{R}$	Percent of data set with $\widehat{R} \leq 1.01$	97.5th percentile of $\widehat{R}$
Quantal-Linear	7,613 (1,758, 12,277)	99.8	1.0011	100	1.0008
Logistic	3,310 (220, 7,632)	97.9	1.0075	99.8	1.0036
Probit	3,423 (68, 7,274)	97.1	1.0106	97.7	1.0084
Weibull	2,054 (379, 8,233)	99.2	1.0051	99.2	1.0056
Multistage 2	5,104 (491, 9,346)	99.4	1.0026	99.2	1.0030
LogLogistic	1,745 (359, 6,687)	99.2	1.0066	99.4	1.0065
LogProbit	1,448 (135, 6,847)	96.1	1.0137	96.3	1.0133
Dich Hill	829 (94, 1,819)	95.2	1.0152	94.4	1.0196
Linear	8,012 (3,515, 13,671)	100	1.001	100	1.0009
Power	2,345 (520, 9,280)	99.1	1.0055	98.1	1.0048
Michaelis-Menten	1,697 (223, 5,378)	99.1	1.0080	98.1	1.0086
Hill	541 (31, 2,198)	76.9	1.0368	75.9	1.0655
Exponential 2	8,048 (4,845, 9,687)	100	1.0009	100	1.0007
Exponential 3	2,159 (519, 9,203)	100	1.0052	99.1	1.0080
Exponential 4	1,440 (6, 8,068)	75	1.1938	79.6	1.3737
Exponential 5	478 (14, 1,653)	75.9	1.1231	73.1	1.1128

Note: CI, confidence interval; MCMC, Markov chain Monte Carlo. The highest $\widehat{R}$ of the parameters in each model is used to calculate the percentage of the data sets with $\widehat{R} \leq 1.01$ and the 97.5th percentile of $\widehat{R}$ for the model.

the tree depth (Stan DevelopmenTeam) achieved and the number of divergent steps. The minimum and maximum tree depth achieved are reported in the model output file included in the results package (see the Supplemental Material, "Testing Datasets and Results" zip file). BBMD uses the default Stan setting for the maximum tree depth allowed (i.e., 10). The results indicate that the Hill model hits the maximum tree depth twice, and the Linear, Power, Michaelis-Menten and Dichotomous Hill model hit the maximum tree depth once. Therefore, we believe that it is appropriate to use the default setting for tree depth. The results show that models with ≥3 parameters have a large number of divergent steps for many testing data sets, which indicates a high risk of obtaining biased estimates. We increased the target acceptance rate "adapt_delta" value (one control parameter for MCMC in Stan) from 0.8 (default) to 0.9 (and even to 0.99 with a step size of 0.01), but the number of divergent steps was only slightly reduced. We further conducted a simulation study to examine the relationship between divergent transitions and bias in BMD estimation (see Supplemental Material, "Simulation Study on the Relationship between Divergence and Bias"; see also Table S1 and Figures S37–S39). The results suggest that the correlation between divergence and bias in BMD estimates is not strong, but the divergence is closely related to some aspects of the dose–response data being modeled (e.g., the number of animals in each group, the number of dose groups, within-dose-group variance). Therefore, in the practice of chemical risk assessment (typically, we only have very limited observations), divergence is to some extent unavoidable for complex models (e.g., the Hill model). A potential way of reducing the number of divergent transitions is employing more informative priors to adequately reduce the space where parameters are sampled, which is our next major task in development of the system.

In addition, we compared the best-fitting model suggestion from BMDS and BBMD. In BMDS, the model with the lowest AIC value is the one suggested for use in BMD analysis, whereas the model with the highest posterior weight should be selected for use in BBMD (if the model-averaging method provided in BBMD is not used). For dichotomous data, 32% (164 of 518) of all data sets select the same best-fitting model based on these criteria between BMDS and BBMD. In BBMD, the Quantal-linear, Probit, and Logistic models (all of which are two-parameter models) are most frequently selected as the best model, whereas BMDS prefers the LogLogistic, Multistage 2, and Quantal-linear models (however, a majority of the selected Multistage 2 model has a reduced format to Quantal-linear). For continuous data, 66 (out of 108) matched the best model, or approximately 61% agreement. In BMDS, the Linear model, the Hill model, and the Exponential 2 model were most likely to be selected as the best model, whereas in BBMD, the Linear, Exponential 2, and Exponential 4 models were the most frequently selected. Interpretation of this comparison is difficult; in BMDS, when model parameter(s) hit a bound and the model is reduced to a simpler format, the AIC value (used for model selection) is calculated based on the reduced number of parameters, which may explain why the Multistage 2 and LogLogistic models frequently have lower AIC values in BMDS.

In testing the BBMD software, we allowed the Dichotomous Hill model, the Hill model, and the Exponential 5 model (each of which has four parameters) to fit the data sets with only three dose groups (which is not allowed in the BMDS) to test the robustness of the BBMD in overloaded conditions. The results presented in Tables 1 and 2 and the dose–response plots included in the results package (see Supplemental Material, "Testing Datasets and Results" zip file) suggest that data sets with three dose groups can be adequately fit by the four-parameter models, but the variance of the posterior sample of model parameters (and further, the variance of the BMD estimates) may be affected. With respect to run time, the mean and 95th percentile interval of running time to fit all eight models to the testing data sets are 16.6 (12.5–46.1) seconds for dichotomous data and 28.2 (11.2–51.9) seconds for continuous data, using default settings. The run time is sensitive to the length of the MCMC chain and the number of chains, although run time estimates may vary depending on hardware.

Test data sets, BMD analysis results on the test data sets for both BMDS and BBMD, and additional BBMD results (including model-specific results on PPP value, BMD/BMDL estimates, model parameter estimates, and convergence of MCMC sampling) are provided in the results package (see Supplemental Material, "Testing Datasets and Results" zip file). Files included in the "Testing Datasets and Results" zip file are listed and described in Appendix 2.

Discussion

Compared with BMDS, the BBMD system generally provides fewer failed BMD and BMDL estimates. The Hill model failed for BBMD in only one case. This failure is primarily due to the plateau feature of the Hill model, that is to say, the shape of the Hill curve in the high-dose range may reach a response plateau; in other words, the response may reach a maximum value. If the

maximum response is smaller than the specified BMR (e.g., a 10% increase of the control response), then the BMD cannot be estimated given the BMR as 10% relative change. This failure can be avoided by changing the BMR to a smaller number (e.g., 1% relative change). In BMDS, the same data set did not provide BMD (or BMDL) estimates owing to an insufficient number of dose groups for the Hill model. Except for this single data set failure, the BBMD system successfully estimated the BMD and the BMDL for all testing data sets, demonstrating the robustness of the BBMD system.

With respect to the BMDL estimation, BBMD overall generated smaller BMD/BMDL ratios for most models and data sets compared with BMDS. For some of the simpler models (including the Quantal-linear, Linear, and Exponential 2 models), the two systems were in good accord on the ratios. However, differences were observed with a smaller BMR of 1% for the Logistic and Probit models. BMDS generated much higher BMD/BMDL ratios when compared with other two-parameter models, and it generated some errors in BMDL calculations for the Logistic model and the Linear model. For models with three or more parameters, BMD/BMDL ratios generated by BMDS were consistently larger than their counterparts calculated by BBMD (with the exception of the Dichotomous Hill model, as described later), including some extreme values in the Hill model, even though those models may have been reduced to a simpler form in BMDS. These high BMD/BMDL ratios may be related to the disadvantage of the profile likelihood method in BMDL estimation described by Moerbeek et al. (2004). The Dichotomous Hill model was the only model that had higher BMD/BMDL ratios in BBMD than in BMDS. This may be because *a*) there were 186 additional data sets that were fit in BBMD but not in BMDS (these data sets had three dose groups and therefore could not be fit in BMDS); and *b*) the Dichotomous Hill model was reduced to a simpler form in BMDS in 124 out of the remaining 332 data sets. Although the convergence indicator, $\widehat{R}$, and the dose–response plots in the results package (see Supplemental Material, "Testing Datasets and Results" zip file) show that BBMD can plausibly fit the four-parameter Dichotomous Hill model to three-dose data sets, this finding suggests that users should give extra attention to the BMDL generated by BBMD in such situations.

PCCs were calculated to examine the similarity between model estimates from the two systems. The PCC was very close to 1 for the two-parameter models, and higher than 0.8 for almost all of the models with three or four parameters. Similarly, the ratio of BMD estimates (BBMD estimates over BMDS estimates) and the ratio of BMDL estimates for the two-parameter models varied around 1. Both pieces of evidence indicate that the two-parameter models perform similarly in both systems. For other models, the median values of the ratios were all within the range of 0.5 to 2, and most were between 0.8 and 1.6. Except for some extremely large ratios in the Hill, Exponential 4, and Exponential 5 models that were mainly caused by the very low BMDL estimates generated in BMDS, most ratios were within the range of 30-fold. We also used linear regression plots to graphically show some outliers of these estimates (see Figures S7–S36). It is worth noting that the estimated parameters in the linear regression should be carefully used to judge equality between the estimates from the two systems because they are very sensitive to outliers. Large differences in BMD or BMDL estimates can have many causes, including the automatic model format simplification in BMDS and the difference between the profile likelihood estimation method and the MCMC approach for BMDL estimation. Model reduction in BMDS was observed frequently in the Weibull, LogLogistic, LogProbit, and Dichotomous Hill models in BMDS because they tend to exactly fit the response rate at the control dose group if it is 0. Overall, 226 (out of 518) Weibull models, 234 (out of 518) LogLogistic models, 214 (out of 518) LogProbit models, and 144 (out of 332) Dichotomous Hill models had a background parameter estimated to be exactly 0. To enforce fitting the curve to pass the 0 response at control can make the BMD estimate either high or low depending on the shape of the curve determined by the remaining dose groups. In contrast, an exact fit at 0 cannot occur in the MCMC sampling process because of its probabilistic nature. Additionally, in BBMD, a uniform distribution between 0 and 1 is employed as the prior for the background parameter (i.e., parameter "a" of the models in Appendix 1), so that it is not informative for the MCMC sampling to search for the optimal range for the parameter. Given these differences in the settings and nature of the methods between BBMD and BMDS, dose–response models in the two systems may show quite different performance in the low-dose range, causing a difference of more than one order of magnitude in the BMD estimates. A simulation study will be conducted to more comprehensively evaluate the two systems on BMD estimation.

The results summarized in Table 3 suggest that simply increasing the number and length of the chains may not necessarily guarantee better convergence. In addition, an $\widehat{R}$ value >1.01 or even >1.05 should not disqualify the posterior sample from being used in a BMD analysis. It is possible that, owing to the features of a particular data set, not all $\widehat{R}$ values for a dose–response model can be <1.01. Therefore, it is important to use multiple criteria (such as visual inspection of the dose–response plot) to judge if the posterior sample can be properly used.

In addition to the uniform priors, we also tested using normal distributions with large variance [i.e., $N \sim (0, 1{,}000^2)$] as priors for model parameters. These normal distributions were truncated so that they had the same lower and upper bound as the uniform priors. The results show that the parameter and BMD estimates from these two different sets of priors were nearly identical. The purpose of using another set of flat priors was not to propose another option for priors but to verify that the uniform priors in the system were appropriate flat priors and not sensitive to the distribution type. A key advantage of Bayesian methods over frequentist methods is the incorporation of prior information on model parameters and model weight. Using adequate informative priors is challenging but provides a great opportunity to utilize the prior information to make the dose–response analysis more reliable. In the example presented in "The Impact of Generalized Informative Prior on BMD Estimation" in the Supplemental Material (see also Figures S3–S6), we used the Loglogistic model to show how an informative prior derived from real toxicological data may affect model fitting and BMD estimation. It is important to note that different priors can result in different inferences.

Rarely, the BBMD system may fail during MCMC sampling in an analysis; in some cases, changing the random seed used in sampling is one possible solution. However, a more fundamental solution is to use more appropriate prior distribution and/or initial values for model parameters, which is our next major task in development. We also noted that some data sets can be successfully analyzed by the BBMD system but not by the BMDS. For example, the Multistage 2 model in the BMDS failed to properly fit data sets number 35 and number 36 in the dichotomous data (see Testing Data_Dichotomous.csv in the "Testing Datasets and Results" zip file), but the Multistage 2 curves fitted in BBMD seem very reasonable.

Conclusion

BBMD is a Bayesian and probabilistic benchmark dose modeling software system with many advanced features for BMD

estimation, including functionalities to estimate model-averaged BMD and "hybrid" approach BMD. BBMD can provide probabilistic estimates for important quantities of interest in dose–response assessment, which greatly facilitates the current need for conducting probabilistic risk assessment. In the next phase of system development, we will conduct research on eliciting informative model parameter priors so that more appropriate priors can be implemented to increase the reliability and robustness of the system. Additionally, the outcomes from the system can be directly used for computing the probabilistic reference dose (or "target human dose") under the framework proposed by Chiu and Slob (2015) and the International Programme on Chemical Safety (IPCS) (2014). Other future areas of research and development include simulating low-dose extrapolation via Monte Carlo simulation.

Appendix 1

Dose–response models for dichotomous data:

1. Quantal-linear model: $f(d)=a+(1-a)\times[1-\exp(-b\times d)]$, $0\le a\le 1,\ b\ge 0$
2. Probit model: $f(d)=\Phi(a+b\times d),\ b\ge 0$
3. Logistic model: $f(d)=1/[1+\exp(-a-b\times d)],\ b\ge 0$
4. Weibull model: $f(d)=a+(1-a)\times[1-\exp(-b\times d^{g})]$, $0\le a\le 1,\ b\ge 0,\ g\ge restriction$
5. Multistage (2nd degree) model: $f(d)=a+(1-a)\times[1-\exp(-b\times d-c\times d^{2})],\ 0\le a\le 1,\ b\ge 0, c\ge 0$
6. LogLogistic model: $f(d)=a+\left((1-a)/\left\{1+\exp[-b-g\times\log(d)]\right\}\right), 0\le a\le 1,\ g\ge restriction$
7. LogProbit model: $f(d)=a+(1-a)\times\Phi[b+g\times\log(d)]$, $0\le a\le 1,\ g\ge restriction$
8. Dichotomous Hill model: $f(d)=a\times b+\left((a-a\times b)/\left\{1+\exp[-c-g\times\log(d)]\right\}\right), 0<a\le 1, 0<b<1,\ g\ge restriction$

Dose–response models for continuous data:

1. Linear model: $f(d)=a+b\times d,\ a>0$
2. Power model: $f(d)=a+b\times d^{g},\ a>0,\ g\ge restriction$
3. Michaelis-Menten model: $f(d)=a+\left[(b\times d)/(c+d)\right], a>0, c>0$
4. Hill model: $f(d)=a+\left[(b\times d^{g})/(c^{g}+d^{g})\right],\ a>0,\ c>0,\ g\ge restriction$
5. Exponential 2 model: $f(d)=a\times\exp(b\times d),\ a>0$
6. Exponential 3 model: $f(d)=a\times\exp(b\times d^{g}),\ a>0,\ g\ge restriction$
7. Exponential 4 model: $f(d)=a\times[c-(c-1)\times\exp(-b\times d)]$, $a>0,\ b>0,\ c>0$
8. Exponential 5 model: $f(d)=a\times[c-(c-1)\times\exp(-(b\times d)^{g})], a>0, b>0, c>0, g\ge restriction$

Note: parameter "a" in the dose–response models above generally represents the response at background dose level, and parameter "g" is a power parameter on the dose. Parameter "b" is the potency parameter in most cases, but parameter "c" may have different meanings in different models; "d" represents dose, which is an independent variable in these dose-response models.

Appendix 2

The detailed analysis results included in the results package zip file contain:

1. "Testing Data_Continuous.csv" and "Testing Data_Dichotomous.csv" contain testing data sets for continuous data and dichotomous data, respectively;
2. "BBMD_Results_Continuous Data.csv" and "BBMD_Results_Dichotomous Data.csv" contain BBMD-generated BMD estimates from eight continuous models and eight dichotomous models, respectively. For each model, the PPP value, model weight, and BMD and BMDL values based on various BMD definitions are reported.
3. "BMDS_Results_Continuous Data.csv" and "BMDS_Results_Dichotomous Data.csv" contain BMDS generated BMD estimates from seven continuous models and eight dichotomous models, respectively. For each model, the *p*-value, the AIC value, and BMD and BMDL values based on various BMD definitions are reported.
4. Sixteen zip files are included. Each zip file contains one csv file (which reports PPP value, BMD and BMDL estimates, model parameter estimates, and convergence indicator $\widehat{R}$ of MCMC sampling, effective sample size, minimum tree depth achieved, maximum tree depth achieved, maximum tree depth allowed, and number of divergence steps), and dose–response plots for all testing data sets.

Acknowledgements

This research was supported by Indiana University School of Public Health Developmental Research Grants for Pre-Tenure Faculty. We also would like to thank J. Gift and M. Wheeler for their comments on an earlier version of the manuscript and Q. Chen and Z. Zhou for their help with the data analyses using Benchmark Dose Software (BMDS).

References

Axelrad DA, Baetcke K, Dockins C, Griffiths CW, Hill RN, Murphy PA, et al. 2005. Risk assessment for benefits analysis: Framework for analysis of a thyroid-disrupting chemical. J Toxicol Environ Health Part A 68 (11-12):837–855, PMID: 16020180, https://doi.org/10.1080/15287390590912153.

Carpenter B, Gelman A, Hoffman M, Lee D, Goodrich B, Betancourt M, et al. 2017. Stan: a probabilistic programming language. J Stat Soft 76(1), https://doi.org/10.18637/jss.v076.i01.

Chiu WA, Slob W. 2015. A unified probabilistic framework for dose–response assessment of human health effects. Environ Health Perspect 123(12):1241–1254, PMID: 26006063, https://doi.org/10.1289/ehp.1409385.

Crump KS. 1995. Calculation of benchmark doses from continuous data. Risk Analysis 15(1):79–89, https://doi.org/10.1111/j.1539-6924.1995.tb00095.x.

EFSA Scientific Committee, Hardy A, Benford D, Halldorsson T, Jeger MJ, Knutsen KH, More S, et al. 2017. Update: Guidance on the use of the benchmark dose approach in risk assessment. EFSA J 15(1):e04658, https://doi.org/10.2903/j.efsa.2017.4658.

Evans JS, Rhomberg LR, Williams PL, Wilson AM, Baird S. 2001. Reproductive and developmental risks from ethylene oxide: A probabilistic characterization of possible regulatory thresholds. Risk Anal 21(4):697–717, PMID: 11726021, https://doi.org/10.1111/0272-4332.214144.

Fang Q, Piegorsch WW, Barnes KY. 2015. Bayesian benchmark dose analysis. Environmetrics 26(5):373–382, https://doi.org/10.1002/env.2339.

Gaylor DW, Kodell RL, Chen JJ, Krewski D. 1999. A unified approach to risk assessment for cancer and noncancer endpoints based on benchmark doses and uncertainty/safety factors. Regul Toxicol Pharmacol 29 (2 Pt 1):151–157, PMID: 10341145, https://doi.org/10.1006/rtph.1998.1279.

Gelman A, Carlin JB, Stern HS, Rubin DB. 2004. *Bayesian Data Analysis. 2nd ed.* Boca Raton, FL:Chapman and Hall/CRC Press.

Hattis D, Baird S, Goble R. 2002. A straw man proposal for a quantitative definition of the RfD. Drug Chem Toxicol 25(4):403–436, PMID: 12378950, https://doi.org/10.1081/DCT-120014793.

Hoeting JA, Madigan D, Raftery AE, Volinsky CT. 1999. Bayesian model averaging: a tutorial. Stat Sci 14(4):382–417.

IPCS (International Programme on Chemical Safety). 2014. *Uncertainty in hazard assessment.* Geneva, Switzerland: World Health Organization. http://www.who.int/ipcs/methods/harmonization/areas/hazard_assessment/en/ [accessed 29 August 2016].

Kass RE, Raftery AE. 1995. Bayes factors. J Am Stat Assoc 90(430):773–795, https://doi.org/10.1080/01621459.1995.10476572.

Millman KJ, Aivazis M. 2011. Python for scientists and engineers. Comput Sci Eng 13(2):9–12, https://doi.org/10.1109/MCSE.2011.36.

Moerbeek M, Piersma AH, Slob W. 2004. A comparison of three methods for calculating confidence intervals for the benchmark dose. Risk Anal 24(1):31–40, PMID: 15027998, https://doi.org/10.1111/j.0272-4332.2004.00409.x.

NRC (National Research Council). 2009. *Science and Decisions: Advancing Risk Assessment.* Washington, DC:The National Academies Press.

Oliphant TE. 2007. Python for Scientific Computing. Comput Sci Eng 9(3):10–20, https://doi.org/10.1109/MCSE.2007.58.
Shao K, Small MJ. 2011. Potential uncertainty reduction in model-averaged benchmark dose estimates informed by an additional dose study. Risk Anal 31(10):1561–1575, PMID: 21388425, https://doi.org/10.1111/j.1539-6924.2011.01595.x.
Shao K, Gift JS, Setzer RW. 2013. Is the assumption of normality or log-normality for continuous response data critical for benchmark dose estimation? Toxicol Appl Pharmacol 272(3):767–779, PMID: 23954464, https://doi.org/10.1016/j.taap.2013.08.006.
Shao K, Gift JS. 2014. Model uncertainty and Bayesian model averaged benchmark dose estimation for continuous data. Risk Anal 34(1):101–120, PMID: 23758102, https://doi.org/10.1111/risa.12078.
Simmons SJ, Chen C, Li X, Wang Y, Piegorsch WW, Fang Q, et al. 2015. Bayesian model averaging for benchmark dose estimation. Environ Ecol Stat 22(1):5–16, https://doi.org/10.1007/s10651-014-0285-4.
Slob W, Setzer RW. 2014. Shape and steepness of toxicological dose–response relationships of continuous endpoints. Crit Rev Toxicol 44(3):270–297, PMID: 24252121, https://doi.org/10.3109/10408444.2013.853726.
Slob W. 2002. Dose–response modeling of continuous endpoints. Toxicol Sci 66(2):298–312, PMID: 11896297, https://doi.org/10.1093/toxsci/66.2.298.
U.S. EPA (Environmental Protection Agency). 2012. "Benchmark Dose Technical Guidance." EPA/100/R-12/001. Washington, DC:Risk Assessment Forum, U.S. Environmental Protection Agency.
Wasserman L. 2000. Bayesian model selection and model averaging. J Math Psychol 44(1):92–107, PMID: 10733859, https://doi.org/10.1006/jmps.1999.1278.
Wheeler MW, Bailer AJ. 2007. Properties of model-averaged BMDLs: A study of model averaging in dichotomous response risk estimation. Risk Anal 27(3):659–670, PMID: 17640214, https://doi.org/10.1111/j.1539-6924.2007.00920.x.
Wheeler MW, Cole T, Bailer AJ, Park B, Shao K. 2017. Bayesian quantile impairment threshold benchmark dose estimation for continuous endpoints. Risk Anal 37(11):2107–2118, PMID: 28555874, https://doi.org/10.1111/risa.12762.
Wheeler MW, Shao K, Bailer AJ. 2015. Quantile benchmark dose estimation for continuous endpoints. Environmetrics 26(5):363–372, https://doi.org/10.1002/env.2342.
Wignall JA, Shapiro AJ, Wright FA, Woodruff TJ, Chiu WA, Guyton KZ, et al. 2014. Standardizing benchmark dose calculations to improve science-based decisions in human health assessments. Environ Health Perspect 122(5):499–505, PMID: 24569956, https://doi.org/10.1289/ehp.1307539.
Woodruff TJ, Wells EM, Holt EW, Burgin DE, Axelrad DA. 2007. Estimating risk from ambient concentrations of acrolein across the United States. Environ Health Perspect 115(3):410–415, PMID: 17431491, https://doi.org/10.1289/ehp.9467.